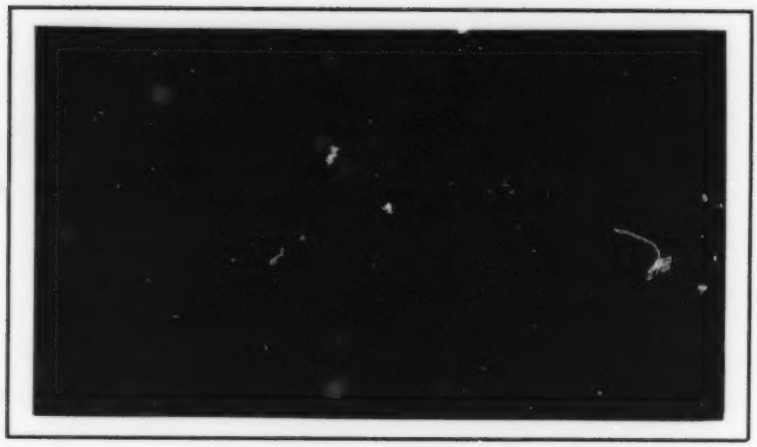

CENTRE FOR
SYSTEMS
SCIENCE

Simon Fraser University, Burnaby, B.C., Canada V5A 1S6 / (604) 291-4704

92-05403/

Compositionality and Systematicity in Connectionist Language Learning

Robert F. Hadley

CSS-IS TR 92-03

February 1992

Compositionality and Systematicity in Connectionist Language Learning

Robert F. Hadley

School of Computing Science
Simon Fraser University
Burnaby, B.C., V5A 1S6 Canada
hadley@cs.sfu.ca

ABSTRACT

In a now famous paper, Fodor and Pylyshyn (1988) argue that connectionist networks, *as they are commonly constructed and trained*, are incapable of displaying certain crucial characteristics of human thought and language. These include the capacity to employ *compositionally structured representations* and to exhibit *systematicity* in thought and language production. Since the appearance of Fodor and Pylyshyn's paper, an number of connectionists have produced what purport to be *counter-examples* to the Fodor-Pylyshyn thesis. The present work examines several of these purported counter-examples. It is argued that although these networks appear to discover a *degree* of compositionality, and to display a degree of systematic behaviour, the degrees involved are substantially less than that found in humans, and (consequently) are less than what Fodor & Pylyshyn require (or presumably would require if the question were put to them).

1. Introduction

In a now famous paper, Fodor and Pylyshyn (1988) argue that connectionist networks, *as they are commonly constructed and trained*, are incapable of displaying certain crucial characteristics of human thought and language. These include the capacity to employ *compositionally structured representations* and to exhibit *systematicity* in thought and

language production.[1] Since the appearance of Fodor and Pylyshyn's paper, an number of connectionists have produced what they believe to be *counter-examples* to the Fodor-Pylyshyn thesis. In the present work I examine six of these purported counter-examples; these are due to Elman (1990), St. John and McClelland (1990), Chalmers (1990), Smart (1990), Pollack (1990) and Smolensky (1990). As will emerge, I argue that although most of these networks appear to discover a *degree* of compositionality, and to display a degree of systematic behaviour, the degrees involved are substantially less than that found in humans, and (consequently) are less than what Fodor & Pylyshyn require (or presumably would require if the question were put to them).

2. Compositionality and Systematicity

In this section I examine a number of specific experiments which establish, to varying degrees, that c-nets can discover the compositionality implicit in a training corpus of *sentences*. When describing the results of these learning experiments, researchers commonly argue from the fact that a network can correctly process *novel* sentences (not contained in the training corpus) to the conclusion that the network has indeed induced a compositional structure, and as a consequence is able to exhibit a degree of *systematicity*. As we examine the c-net experiments described below, it will be useful to distinguish different degrees of systematicity, according to the *degree of novelty* of sentences which a c-net is able to recognize (given the c-net's training regime). I shall distinguish three degrees of systematicity. No doubt, it would be possible to make even finer distinctions, but for our purposes the following should suffice. The degrees of systematicity are: weak, quasi, and strong.

1) *Weak Systematicity.* Networks exhibiting weak systematicity can perform at least the following kind of generalization: Suppose that a training corpus is "representative" in the sense that every *word* (noun, verb, etc.) that occurs in some sentence of the corpus also occurs (at some point in the training corpus) in every permissible syntactic position. Thus, although the training corpus omits some sentences permitted by the target grammar, any network trained on this corpus will have been trained to recognize every word in every syntactic position that the word will occupy in the set of novel test sentences which are used to demonstrate the network's generalization capacity. Assuming

[1] 'Compositionality' here presupposes that representations have a combinatorial syntax and semantics, whereas 'systematicity' refers to the systematic relationships which *result* when such combinatorially structured representations are employed.

that this set of novel sentences contains only sentences which are syntactically isomorphic to sentences in the training corpus, and that no new vocabulary is present, we shall say that a c-net exhibits at least *weak systematicity* if it is capable of successfully processing (by recognizing or interpreting) novel test sentences, once the c-net has been trained on a corpus of sentences which are *representative* in the sense described above. I describe such c-nets as (at least) weakly systematic in order to reflect the fact that their generalization capacity has only been tested upon sentences which are *weakly novel* with respect to the training corpus.

2) *Quasi-Systematicity.* We shall say that a system exhibits only quasi-systematicity if (a) the system *can* exhibit weak systematicity, (b) the system successfully processes novel sentences containing embedded sentences, such that both the larger containing sentence and the embedded sentence are (respectively) structurally isomorphic to various sentences in the training corpus, (c) for each successfully processed novel sentence containing a word in an embedded sentence (e.g., 'Bob knows that Mary saw *Tom*') there exists some *simple* sentence in the training corpus which contains that same word in the same syntactic position as it occurs within the embbeded sentence (e.g., 'Jane saw *Tom*'). A system would be *merely* quasi-systematic if 'Tom' needed to occur (in the training corpus) in the *object position* of a *simple sentence*, before the system could correctly process embedded occurrences of 'Tom' in object position. Analogous remarks apply to subject position, verb position, etc.

3) *Strong Systematicity.* We shall describe a system as strongly systematic if (i) it *can* exhibit weak systematicity, (ii) it can correctly process *simple* novel sentences containing words in positions where they *do not appear* in the training corpus (i.e., the word within the novel sentence does not appear *in that same syntactic position* within any *simple or embedded* sentence in the training corpus). Note that a system which has not been trained on embedded clauses may still exhibit strong systematicity, because neither condition (i) or (ii) requires that embedded sentences be present.

Now that I have distinguished various *degrees* of systematicity, we may examine several recent attempts at connectionist language acquisition, and consider the degrees to which these various attempts learn compositional structure and exhibit systematicity. Before doing so, however, I shall argue that *humans* exhibit the strongest of the three forms of systematicity described above. To begin with, let us note that even young children, who have not yet reached the stage of producing multi-word utterances, are frequently able to

obey *simple* imperative sentences which contain words in syntactic positions where the child has never encountered the word before.

It is well known, for example, that in the few weeks which precede a child's first multi-word utterances, a "spurt" occurs in a child's acquisition of *nominals* (both common and proper nouns), and that during this period children are able *rapidly* to acquire the use of nominals by means of "what's that" games (cf. (Ingram, 1989; Dromi, 1987)). Once they have acquired nominals in this fashion, children are soon thereafter (i.e., within minutes) able to comprehend these words in sentences they encounter. This fact is established by Katz, Baker, and Macnamara (1974) who also present a strong case that the ability of young children to distinguish proper nouns from common nouns is much more a function of a child's *prior* ability to distinguish re-identifiable individuals from classes of objects than it is a function of some capacity to distinguish words which are syntactically preceded by an *article* from those which are not.

Moreover, as children begin to produce simple, multi-word utterances, they will often produce semantically reasonable, *albeit non-grammatical* combinations of the words they have acquired in previous contexts. In fact, children do this sufficiently often that some psycholinguists posit the existence of a "child grammar" (Ingram, 1989). Now, whether or not we accept the existence of a child grammar, the fact that children are able to recombine words in patterns that are not present in their training corpus strongly suggests that (at least in the early weeks of multi-word utterance) children have a much greater grasp of the *semantic* content of particular words than they do of their syntactic roles (in adult grammars). (The results of Katz, Baker, and Macnamara also reinforce this conclusion.) Furthermore, and more to the point, the ability of children to *sensibly* recombine words in patterns they have not been trained to produce clearly demonstrates that children are not nearly as dependent upon syntactic context as systems which are only weakly (or quasi) systematic.

As we consider somewhat older children, who have acquired a rudimentary syntax (but not necessarily the use of prepositional phrases or relative clauses), it becomes transparently clear that humans can learn to use nominals long before they have encountered them in all possible positions. For example, a child visiting a zoo with her parents may hear her mother exclaim, "Susie, look at the otter". Susie may reply, "What's an otter?" The mother, pointing, replies "Here, this is an otter". If Susie is adept at language, she may learn the (approximate) meaning of 'otter' rapidly, by this ostensive means, and may

soon utter, "Look, Mommy, this otter is chasing the other one". Although the child has never encountered the word in *subject* position, she is able to use it in that position once its meaning has been surmised. Of course, most children will require a few repetitions before an ostensively introduced word enters long term memory, but these repetitions need not present the word in all legal positions. With adults, new words may enter the vocabulary even more rapidly, as when one surmises a word's meaning during the course of conversation, or when listening to a brief exchange during a meeting. Once a word's meaning has been surmised, most adults can use it freely in embedded sentences and simple sentences, although they may only have heard the word used in a single syntactic position.

- - - - - - - - - - - - - - - - -

We turn now to consider connectionist systems which, *prima facie*, challenge Fodor and Pylyshyn's view on the limitations of c-nets vis-a-vis compositionality and systematicity. In considering these systems we should bear in mind that Fodor and Pylyshyn are concerned with the kind of full-fledged compositionality and systematicity that human thought and language exhibit.

2.1 St. John and McClelland

St. John and McClelland (1990) present a connectionist model which learns to assign "semantic representations" to English sentences which are presented as input. Although the details of their model are somewhat complex, the overall gist is that, via backpropagation, the network is trained to produce a correct semantic representation of the situation *described* by each input sentence. Situations (or events) described by input sentences consist of relationships, and the objects involved in those relationships. Input sentences are fed into the network in presegmented constituents. As each constituent is processed, an inspection is made to see whether the network has output the desired, complete representation of the target situation. Backpropagation is performed after each such inspection. Because it is usually not possible to predict the *entire* target representation on the basis of isolated sentence constituents, the network is forced to learn associations between individual constituents and particular objects or relations in the target situation.

Sentences which serve as input constitute a highly simplified version of English, in that all articles are deleted and only singular nouns are present. However, certain preposi-

tional phrases are permitted. Each target semantic representation consists of an *ordered* series of role/filler pairs. Roles are *agent, action, patient,* etc., and fillers are "concepts" (*my* scare quotes) corresponding to individual nouns and verbs. Thus, each semantic representation is a *structured, concatenated* sequence of pairs. By itself, this aspect of the model would seem to undermine any potential the model might possess for deposing Fodor and Pylyshyn's thesis that human thought requires structured, internal representations. For the experimental design presupposes the existence of such representations (at the point where backpropagation is employed). Moreover, the ability to form such representations *presupposes* that the learner has *already discovered* a compositional, systematic method of representing situations. Thus, from Fodor and Pylyshyn's standpoint, the model's design concedes one of their major contentions. However, the question still remains whether the model acquires knowledge of the compositional, systematic nature of its *input* sentences. St. John and McClelland (hereafter, St.J&Mc) clearly claim that it does (p. 250, 1990), and it is this claim we now consider.

As mentioned, St.J&Mc's training corpus includes sentences containing prepositional phrases. Unfortunately, when testing their network for the acquisition of compositional knowledge (which is manifested as *systematicity*) St.J.&Mc used simpler training corpora, which lacked prepositional phrases. Two experiments were conducted to test for systematicity of behaviour – one syntactic, the other semantic. Tests for syntactic systematicity involved only 10 objects and 10 reversible actions. Each object (action) uniquely corresponds to a particular noun (verb) in the training corpus. Both active and passive verb forms were permitted, and each input sentence had the general form: [noun verb-form noun]. Given that both active and passive forms are possible, a total of 2000 sentences are possible. All 2000 sentences were generated. Of these, 1750 comprised the training corpus, and the remaining 250 were set aside for later testing. Although St.J&Mc do not explicitly say so, their remarks elsewhere (p. 243, 1990) suggest that these 250 sentences were randomly selected. Assuming they were, it is highly probable that the remaining 1750 sentences contained occurrences of every word in every legal syntactic position. (Otherwise, 80% of the 250 test sentences would have to contain the *same* particular noun or verb in the same syntactic position. Given that there are 10 nouns and 10 verbs, this is extremely unlikely.[2]) Moreover, St.J&Mc give *no indication* that

[2]Note that 10% of the 2000 original sentences contain a given noun or verb in a given position. So, if a given noun or verb does not occur in a given position within the 1750 training sentences, then 200 of the 250 test sentences must contain that given word in the given position.

the training corpus (of 1750 sentences) *does not* include every possible word in every possible position. On the available evidence, therefore, it is reasonable to believe that the training corpus does present every word in every possible position. This conclusion is reinforced by St.J&Mc's remark that "What makes this a generalization task is that some of the sentences were set aside and not trained: *some agents were never paired with certain objects*" (my emphasis). The fact that sentences in the *test* corpus describe novel agent-object combinations does present convincing evidence of generalization, but does not suggest that anything stronger than weak systematicity and compositionality were tested for. To be sure, the network does display some degree of systematicity. The network assigns the correct semantic representation to 97% of the novel 250 sentences. However, given the above considerations, it seems entirely likely that the network displays only *weak syntactic* systematicity.

The test for *semantic* generalization is analogous, in relevant respects, to the one just described. The semantic test involved a set of 400 possible sentences, of which 350 were used for training and the remaining 50 were used for testing. St.J&Mc explicitly note that the 50 test sentences were randomly selected from the set of 400. As before, the set of 400 sentences exhausts the space of possible sentences. Now, since the 50 test sentences were randomly chosen, it is extremely probable (by analogy with the reasoning given in the previous footnote) that each word occurred in a syntactic position within the *test* corpus that it also occupied within the training corpus. Thus, it is virtually certain that the test for *semantic* generalization established only weak systematicity. Certainly, we are given no reason to suppose otherwise. Also, it is clear that St.J&Mc's model was not even intended to display the kind of strongly systematic behaviour and rapid integration of semantic knowledge which our example of the child at the zoo illustrates (involving the word 'otter').

It should be acknowledged, however, that despite the weaknesses mentioned above, the network we have considered yields some impressive results, including the ability to learn "to disambiguate ambiguous words; instantiate vague words; assign thematic roles; and immediately adjust its interpretation as each constitutent is processed" (p. 220, 1990). Even the ability to demonstrate weak systematicity is no small feat. However, it should be remembered that humans appear to exhibit a much stronger form of systematicity than this. Also, the weak *semantic* compositionality displayed by St.J&Mc's network raises further doubts (as I now argue).

For example, St.J&Mc report that the net has difficulty forming *semantic representations* for sentences reporting radically novel combinations of objects and relations. The example they describe involves two groups of agents who watch different kinds of television programs. In the network's training, children always watch children's shows and adults always watch other shows (e.g., the news). Once the network has been trained, it is fed the sentence 'Bobby watched (the) news", which describes a situation dramatically novel for the network. In cases such as these, the network is unable to assign any coherent interpretation to the sentence.

Now, perhaps it could be argued that "Bobby watched the news" is a *semantically anomolous* sentence for the network, and that humans also have difficulty interpreting semantic anomolies, such as "Bobby ate justice for supper". However, this rejoinder falters on two accounts. First, the sentence "Bobby watched the news" should *not* be considered semantically anomolous, even for the network. For, unlike "Bobby ate justice . . . ", the network has processed many sentences of the form "x watched the news", and many other sentences of the form "Bobby watched x". Thus, the network has learned that 'Bobby' can be the subject of 'watch' and that 'news' can be the object. In cases such as these, humans are able to comprehend sentences reporting dramatically novel situations, although they may experience considerable surprise.[3].

The second difficulty with this rejoinder is that humans are frequently able to assign at least *some* interpretation to sentences such as "Bobby ate justice". It is not as if the individual words in the sentence lose their meaning, for example. Moreover, because humans are so good at composing compound meanings, even when no *mental image* is possible, some linguists and philosophers have questioned whether any sentence is *truly* semantically anomolous, and whether "category mistakes" truly exist.

2.2 Elman

We turn now to the work of Elman (1989, 1990) on connectionist learning of syntactic structure. As indicated in the introduction, Elman contends that "the sensitivity to context which is characteristic of many connectionist models, and which is built-in to the architecture of the networks used here, does not preclude the ability to capture generalizations which are at a higher level of abstraction." In addition, Elman clearly

[3]There is substantial evidence that even 18 month-old children can comprehend commands involving entirely novel verb-noun combinations (cf. Ingram, 1989, p. 166)

opposes his results (and those of others, including St.J&Mc) to the conclusions advanced by Fodor and Pylyshyn (1988), and to Fodor's (1976) Language of Thought thesis. Yet, while it is clear that Elman's networks do generalize and acquire a degree of systematicity, it is by no means clear that they display the degree of systematicity that humans exhibit. Moreover, since Elman's research does not address issues of *semantic* systematicity and compositionality, it is unclear whether this work actually threatens Fodor's views on the Language of Thought. After all, we saw that St.J&Mc were able to train their network to discover *semantic* compositionality only when they assumed the prior existence of a concatenative, structured set of internal representations. However, let us consider Elman's results in some detail.

Elman (1989) describes two experiments, both employing recurrent networks with a context layer feeding back into the hidden layer. The training procedure for both networks is essentially the same. Simplified English sentences (articles are absent) are fed into the network one word at a time, and backpropagation is used in a (*prima facie*) attempt to train the network to *predict* the next word it will receive as input. However, since a large training corpus is employed (10,000 sentences in each experiment), the network cannot learn to predict the next input word, but does learn (in essence) to predict the *syntactic category* of the following word. The first of the two experiments is designed, in fact, to demonstrate that the network does indeed develop a set of syntactic categories which correspond to the traditional grammatical categories. Cluster analysis on the network's hidden-layer activation values reveals that the network acquires *approximately* traditional categories, as well as (approximate) subcategories corresponding to animate noun, inanimate noun, transitive verb, etc.[4] The syntactic corpus for this experiment consists entirely of simple 2 and 3 word sentences. Both singular and plural nouns are included, and the network does learn to detect number agreement.

The second experiment is designed to test whether a somewhat more complex recurrent network can discover syntactic structure. In this experiment the training corpus includes relative clauses, embedded to a maximum depth of two (judging by examples provided). Now, although the acquisition of *approximate* syntactic categories in the first experiment seems to indicate that a *degree* of systematicity has been discovered, only in this latter experiment is a test for systematicity explicitly performed. We therefore concentrate our

[4]These categories only approximate traditional categories because (for example) the representations developed for subject and object tokens of the same noun are not identical, though they do cluster together.

attention upon the latter experiment.[5]

The training regime for the second experiment consisted of four phases, the first of which presented the network with a continuous stream of 10,000 sentences, containing no relative clauses. The three remaining phases each built upon the preceding phases, and involved increasingly high percentages of relative clauses. This controlled, graduated exposure to relative clauses raises questions about the psychological plausibility of the design, which I shall explore in section xxx. However, our present concern is with systematicity. Given that the initial training phase involved 10,000 sentences, comprised only of 8 common nouns, 2 proper nouns, and 12 verbs, we have good reason to suppose that the *initial* training corpus presented every word in every syntactically legal position.[6] Assuming this is so, it is clear that the first phase could induce only weak systematicity. Moreover, there is no reason to suspect that any stronger form of systematicity is established by the later training phases, since every possible arrangement of nouns and verbs that could occur as the complement of a relative clause appears to have been present within simple sentences in the first training corpus.

In passing, it is worth noting that Elman does not say whether his test corpus included greater *depths* of embedding than were present in the training corpus. This is unfortunate, since the ability to generalize to greater depths is an important component of human thought. When only tail recursion is involved (as opposed to center embedding), humans can comprehend surprisingly deeply embedded sentences. For example, consider the sentence: "John knew that every student who had taken a course from the professor who taught the only Kant course which was offered by the Philosophy department came to regret that they had ever taken a course which was offered by the professor who had taught that Kant course."

2.2.1 Fast Integration of New Vocabulary and Semantics

In section 1, I described an example in which a child *rapidly* (i.e, after only a few trials) learns to integrate a novel word, 'otter', into her vocabulary, and to use the word in novel contexts. Examples such as these, which argue for strong systematicity in humans, are

[5]Also, it is quite clear that the training corpus for the first experiment presented every word in every syntactically legal position. This can readily be established on the basis of the number of nouns and verbs available. It follows that the first experiment establishes only weak systematicity at best.

[6]Note that even if we assumed that every verb optionally takes a direct object, the total number of possible *simple* sentences is: [10 nouns × 12 verbs × 10 nouns = 1200] plus [10 nouns × 12 verbs = 120].

problematic for Elman's networks as well as St.J&Mc's. However, in (Elman, 1990) the issue of integrating new vocabulary is briefly addressed. Elman describes a training corpus in which a novel word, 'zog', replaces all occurrences of the word 'man' in a previous training corpus. Although the word 'zog' has its own bit-pattern representation, when the network is trained with the new training corpus a hidden layer representation is developed for 'zog' which is strongly analogous to the previous representation for 'man'. The strong similarity is explained by the identity of surrounding context for 'zog' and 'man'. These results suggest that context may help to speed up the learning of new words to some degree. However, we have no reason to suppose that this speed-up in learning would begin to approach the truly *rapid* acquisition of vocabulary and meaning that we witness in the human case. For, nothing in Elman's discussion would suggest that a network could very rapidly (after 4 or 5 exposures, say) discover that 'zog' could be used in most contexts in which the network has learned that 'man' can be used.

2.3 Chalmers

Chalmers (1990) directly challenges Fodor and Pylyshyn's conclusions concerning the ability of c-nets to discover compositionality and exhibit systematicity. In Chalmer's view, Fodor and Pylyshyn seriously underestimate the potential for c-nets to perform structure mappings between *distributed* representations. He cites the work of Elman (discussed above) and Smolensky (1990, discussed below) as providing strong counterexamples to Fodor and Pylyshyn's claims. In addition, he describes his own experiment in which a network is trained to transform distributed representations of simple sentences in active voice into distributed representations in passive voice. The distributed representations are derived by training a RAAM network (Pollack, 1990) to develop consistent, distributed activation patterns for both active and passive sentences. A separate network is next trained, using backpropagation, to produce the appropriate transformations between active and passive representations (i.e., the distributed patterns). The network is trained using 75 active and 75 passive voice sentence patterns, and within acceptable limits of error, the network learns to output the appropriately transformed activation pattern. In a fashion similar to those examined above, Chalmers tests for acquisition of compositional knowledge by means of a *novel* test corpus of 50 active and 50 passive sentences. However, all words occurring in this novel corpus occur in the initial training corpus, and within this initial corpus all words occur in all syntactically permissable positions (personal communication). Thus, it has only been shown that the network ex-

hibits weak systematicity. While Chalmers establishes a convincing case that c-nets can perform mappings between distributed representations of structured objects, he has not demonstrated that c-nets can exhibit the kind of strong systematicity or compositionality that Fodor and Pylyshyn presumably require.

In passing, it is worth noting that a variant of the "fast-learning" phenomenon arises for the passive-active transformation. Suppose you are asked to give the passive form of "Zim loves Zam". Though you may pause momentarily to wonder about the unusual names, you are able to produce "Zam is loved by Zim" without effort. Now, although context may provide a strong cue that 'Zim' and 'Zam' function as names, c-nets which have been trained to perform passive transformations in the fashion employed by Chalmers and St.J&Mc are certainly not able to produce *immediate* passive transformations upon distributed patterns they have never before encountered. In this connection, it is important that within both St.J&Mc's network, and Chalmer's, the internal distributed representation of 'loves' (after training) is context-sensitive, and that reading the novel word 'Zim' *may* put the network in a *novel* state in which it would not even be able to recognize that 'Zim' and the succeeding 'loves' are separate words. It is significant that humans are able to take an explicit instruction, such as "Give me the passive form of X", and apply it to something which, for the agent, is at least a partially novel representation. I have argued elsewhere that this ability to apply transformations to internal representations *upon demand* (but not as an ordinary practice) suggests that a substantial degree of modularity is at work in behaviour such as this.[7]

2.4 Small

Small (1990) describes a c-net which is trained, via backpropagation, to *predict* the next word in a sentence, given the preceding word as input. Small's training copora contain only 3 or 4 sentences each, and articles are not included in these sentences. Given these limited training sets, the network does learn to predict succeeding words in each sentence. However, in light of the small number of sentences involved, it seems entirely possible that the network has merely memorized particular sequences of data. Unlike Elman (1990), Small presents no evidence that syntactic categories are learned, and the network is not tested with novel sentences of any kind. Thus, it is difficult to see why we should conclude that the network has performed any substantial generalization.

[7]See (Hadley, 1990) for more on the *rapid* application of explicit rules to internal representations.

It is true that Small's results support the claim that a c-net's predictions can be highly dependent upon the preceding sentential context, but *that* claim was not in doubt.

2.5 Pollack

To this point, we have examined four distinct systems, each of which is intended to induce either syntactic or semantic knowledge from quasi-natural input. Thus far, we have found no reason to conclude that any of these systems exhibit strong systematicity. We turn now to the recent work of Pollack (1990) whose Recursive Auto-Associative Memory (RAAM) exhibits a degree of systematicity and productivity within a limited domain. Unlike the systems so far examined, Pollack's RAAM does not extract its knowledge from quasi-natural sentences, but from structured parse trees whose constituents are represented (in Lisp fashion) as nested lists.

As Pollack notes, the architecture of RAAM is essentially that of a simple auto-associative memory (Rumelhart, et al, 1986). The network consists of equal-sized input and output layers, with a smaller hidden layer which develops condensed, distributed representations. Rumelhart, et al, have demonstrated that simple networks of this kind can be trained, via backpropagation, to reproduce symbolic patterns presented to their input layers on their output layers. Merely reproducing patterns on an output layer is, of course, not very interesting. What is interesting is that such networks are able to replicate the input pattern on the basis of information contained in the intermediate, distributed representation. Pollack takes the work of Rumelhart, et al, a few steps further by demonstrating that networks of the above kind can be trained to replicate *nested* structures, such as parse trees, on their output layers. For example, consider the following parse tree which is represented by means of nested lists:

((Det (Adj Noun)) (Verb (Prep (Det Noun))))

A RAAM network can be recursively trained, in stages, to auto-associate a bit-form representation of this tree structure with itself. The training involves *first* training the network to auto-associate on the unnested constituents (e.g., (Adj Noun)), and then on progressively more complex constituents. In the final stage, the network is trained to auto-associate on the entire parse tree.

In one experiment, which demonstrates the productivity and systematicity of RAAM,

the network is trained to auto-associate on 13 *ternary tree* structures, which encode 13 English sentences of increasing complexity. For example, one such ternary tree, which includes an embedded structure, is:

(THOUGHT PAT (KNEW JOHN (LOVED MARY JOHN))).

Once the network is trained on the thirteen ternary trees, it is tested for its ability to auto-associate on *novel* trees, which correspond to *novel* sentences comprised of the same atomic constituents. The RAAM network succeeds in this task, though syntactically mal-formed trees are occasionally generated. Certain of the *successful* trials present noun-verb combinations which are novel in the sense that the particular noun does not appear as an argument to that verb anywhere in the training corpus. Thus, the network has unquestionably induced some of the compositional structure implicit in the training corpus. However, Pollack's sole argument for systematicity rests upon the fact that "All 16 cases of (LOVED X Y), with X and Y chosen from the set JOHN, MARY, PAT, MAN were able to be reliably represented, even though only four of them were in the training set." As it happens, although one of the words, 'man', never appears as argument to 'loves' in the training corpus, each of the four nouns occurs as subject and as direct object somewhere in the training corpus. Thus, *strong* systematicity appears not to have been established. A case for a kind of quasi-systematicity can be made, however, since within the training corpus 'John' occurs as a direct object only within an embedded constituent. Yet, the network successfully processes the novel sentence, (LOVED MARY JOHN), which corresponds to the simple English sentence, 'Mary loved John'.

Furthermore, Pollack's method of bit-encoding nouns and verbs creates the potential for an even stronger form of compositionality, since words which belong to the same *semantic type* are given similar bit-form representations. For example, 'man', 'John', and 'Mary' all have the same corresponding bit (out of 16 bits) set to 1, to indicate the semantic type, 'human'. Bit-typing such as this may be viewed as a shorthand method of expressing the fact that, were the words to have associated semantic content, they would presumably share certain micro-features. In addition, the similarity in bit-form encoding increases the likelihood that words belonging to the same semantic type can be interchanged in similar contexts. However, it appears that semantic typing alone would do little, if anything, to enable c-nets to mimic the human ability to perform *novel* active-passive transformations. Consider the example, given earlier, of being asked to produce

the passive form of 'Zim loves Zam'. Since 'Zim' and 'Zam' cannot be assumed to have *prior* semantic types, our ability to form 'Zam loves Zim' and 'Zam is loved by Zim' does not appear to arise from the semantic aspects of these words. Nor does it seem to derive *primarily* from semantic typing associated with the verb 'loves'. This is evident from the fact that we can readily form the passive of 'The rock loves the sea' even though the semantic types of the nouns clash with that expected by the verb.

A final difficulty with Pollack's approach to the acquisition of linguistic systematicity is that the RAAM network is trained *only* on pre-parsed, explicitly parenthesized structures whose syntactic structure has *already* been analysed. Thus, his training regime, like that of St.J&Mc, is parasitic upon an external agent's having already discovered the systematic properties of the input set. Pollack displays some awareness of this difficulty when he notes, "The simplifying assumption, that internal representations can first be devised and then used as target patterns, is questionable." However, not only is this simplifying assumption questionable, it largely undermines Pollack's case (p. 95, 1990) that RAAM is "still quite systematic, according to Fodor and Pylyshyn's own definition." For, though their definition does not explicitly say so, it is clear from the general thrust of their paper that Fodor and Pylyshyn would expect c-nets to arrive at systematicity under *their own steam.*

2.6 Smolensky

In (Smolensky, 1990), the question whether highly structured, *fully* compositional representations can be realized in completely distributed form is examined from a mathematical perspective. Smolensky's answer is affirmative; he is able to demonstrate with considerable rigour, using tensor calculus as a tool, that concatenative symbolic representations (including representations of nested structures) can be translated into tensor product representations. These representations subsume the vector space representations commonly employed in the analysis of c-nets, and Smolensky demonstrates that *many* existing connectionist systems can straightforwardly be analysed using the tensor calculus.

Smolensky's general strategy is to show (1) that structured symbolic representations can be analysed as *sets* of filler/role pairs, (2) that each such set has a tensor product representation, and (3) there exist systematic ways of instantiating tensor product representations in c-nets, in a fully *distributed* fashion. It is this last point (3) that bears most

directly upon the Fodor-Pylyshyn thesis. For, since the tensor product representations correspond to conventional, structured representations, Smolensky contends that Fodor and Pylyshyn grossly underestimate the power of *distributed* connectionist representations.

Now, although Fodor and Pylyshyn *may* have underestimated the power of distributed representations, it is not clear that Smolensky's results establish that c-nets (as commonly implemented) can exhibit strong systematicity. As we have seen, a system exhibits systematicity when it can *process* its representations in ways that exploit the compositionality of those representations. C-nets can process their representations only if the appropriate *weight* vectors are in place, and Smolensky presents no arguments to show that these weight vectors can be acquired in a manner consistent with the resource limitations of human agents. Moreover, even ignoring the problem of how weight vectors are acquired, the question remains, how are the requisite *distributed* representations learned? For the most part, Smolensky sidesteps this issue, as when he says (p. 193, 1990) "Here it is not the job of the network to set up the tensor product representations: in presenting the input/output pairs to the network during training, the modeler must convert the symbolic inputs and outputs to their vector representations, and this can be done directly by using the mathematical definition of the tensor product representation." Elsewhere, Smolensky explains how to construct the required vector representations from tensor product representations, but what we are told, in effect, is how to hand-construct a series of contrived, grid-like structures whose dimensions are a precise function of the number of fillers and roles involved. No indication is given as to how these grid-like structures might arise through a learning process. Near the end of his paper, Smolensky does briefly discuss how sets of *roles* (which occur in the sets of filler/role pairs which give rise to tensor product representations) might be learned. But there is no indication that the learning processes described would be more powerful than those employed by Elman (whose networks also learn grammatical *roles*), and we have already seen that Elman's method appears to establish only weak systematicity. While it is imaginable that Smolensky's distributed representations of filler/role pairs could be acquired without the kinds of contrived, combinatorially explosive training regimes we have examined thus far, we have no particular reason to suppose this is so.

3. Psychological Plausibility of Training Regimes

In the preceding pages I have occasionally commented upon psychologically problem-

atic aspects of certain of the training regimes invloved. Although none of the models considered here make strong claims for the psychological plausibility of their methods, it is important to consider whether the results obtained actually *require* learning conditions which are truly implausible. For, even *competence* models of cognitive behaviour (as well as performance models) are normally expected to preserve (or at least approximate) extensional relationships between an agent's *real* input and *real* output. If a particular c-net training regime *requires* the existence of input copora or external error feedback which simply do not occur in human conditions, then serious doubts arise as to whether the c-net model can even provide insight into human cognition. This is especially true when there appears to be no way to *modify* the c-nets involved such that more realistic sets of input and output can be accommodated.

In what follows I examine aspects of the work of St.J&Mc (1990) and Elman (1989, 1990) which *prima facie* (at least) involve seriously unrealistic assumptions about certain learning and/or biological conditions involved in human language acquisition. I focus upon St.J&Mc and Elman not because their models are more seriously flawed than the remaining models considered thus far. Rather, it is because the remaining models are not *primarily* concerned with *processes* of language acquisition and are not designed to explain or model the *range* of linguistic phenomena that Elman and St.J&Mc address.

3.1 St. John and McClelland (1990)

Recall that the training regime of St.J&Mc presupposes that the learner has *already* apprehended, at the time a given input *sentence* is processed, the particular external state of affairs that the sentence describes.

The learner apprehends this state of affairs by having a *structured*, sequentially ordered representation of this state of affairs in mind. I have already remarked that these structured representations resemble, in spirit at least, those of Fodor's Language of Thought. However, our present concern is with a different problem, *viz.*, is it legitimate to assume that among all the various states of affairs perceptually available to the agent at the time the sentence is presented, the agent's attention is drawn to the particular state of affairs described by the sentence? For example, imagine that a child is playing on the floor of the living room when her father says, "Bring Mommy her slippers". Now, even assuming that the child has *previously* learned the referent of 'Mommy' (a problem which St.J&Mc ignore in their discussion of bootstrapping), the child may be faced with a bewildering

array of states of affairs from which to choose. Suppose that Mommy is eating an apple, wearing a blue robe, sitting near the fireplace, and watching television. Which of these states of affairs is the child to focus upon? Even if Daddy is attempting to call attention to the slippers is some way, several candidate states of affairs remain. Moreover, it is unrealistic to suppose that *whenever* a child hears a sentence – even simple sentences – someone is striving to fix the child's attention upon the relevant state of affairs.

Now, St.J&Mc briefly address the above difficulty when they say (p. 249, 1990) "The problem of discovering which event in the world a sentence describes when multiple events are present would be handled in a similar way, though we have not modelled it. Again, the aspects of the world that the sentence actually describes would be discovered gradually over repeated trials, while those aspects that spuriously co-occur with these described aspects would wash out". There are two difficulties with St.J&Mc's suggestion. The first is that, given their experimental design, if spurious states of affairs were frequently presented to the backpropagation algorithm as the *intended* target state of affairs, the *number of iterations* required to wash out the spurious information may well be utterly implausible. Even without spurious information, the network requires over 300,000 iterations *before* it begins to master sentences in passive voice. Were a substantial percentage of spurious states of affairs to be presented, the complexity of the learning task would certainly increase, and we have no reason to suppose the number of iterations involved would fall within anything resembling a plausible range. Even the existing figure of roughly 300,000 raises doubts. These doubts would not be so unsettling if we had reason to believe that alternative architectures would dramatically decrease the iterations involved, but the authors present no arguments to that effect. Moreover, we must bear in mind that the learning task has already been dramatically oversimplified by (a) the absence of articles in the input copora and (b) the fact that the agent's internal representation of the target state of affairs contains a *marker* indicating whether the input sentence is in active or passive voice. It is very difficult to see how a *perception* of an external state of the world could yield an indication as to whether the given *sentence* was active or passive. Moreover, it seems implausible that the agent would represent voice information *before* the active-passive distinction had been at least partially mastered, and *that* distinction is found in the relationships between the input sentences and internal representations of states of affairs, not in the latter representations alone.

The second difficulty with St.J&Mc's suggestion that spurious states of affairs would

"wash out" is that their design presupposes that the agent's attention *remains* focused upon a *single* state of affairs while the entire sentence is being processed. Yet, this is a questionable assumption. Given a child's attention span, and the presence of multiple states of affairs, it seems entirely possible that the child's attention will flit from one state of affairs to another during the course of the utterance. This seems especially likely when the given sentence contains prepositional phrases and/or relative clauses.

Related to this difficulty is the way in which St.J&Mc employ the backpropagation algorithm. During the training procedure, as each constituent of a sentence is processed in turn, the resulting pattern on the output layer is compared to the *entire target* state of affairs. Differences are noted, and backpropagation of error is employed *after each constituent is processed* (although weight changes are accumulated and adjusted after each 60 trials). Now the lack of a biological correlate for the standard backpropagation algorithm is a well known problem, but the defense is commonly made that there may exist some unknown biological process whose effects are roughly analogous to those of this algorithm. This defense is reminiscent of the kind of hand-waving that some connectionist find lamentable in classical AI. However, even if this hand-waving response is accepted, and even if we accept a suggestion of Smolensky (1988) that connectionist processing occurs at a more abstract level than the neural level, still there must be *some* biological process which is presumed to support the more abstract process which is supposed to (roughly) correspond to the backpropagation algorithm. Moreover, this biological process would presumably occur each time the backpropagation algorithm is executed in the training regime, and this biological process requires time. Given the complexity of backpropagation, it is difficult to believe that a biological process supporting the algorithm's abstract analogue could occur during the interval between the uttered constituents in a sentence. In light of this, serious doubts arise as to the legitimacy of invoking the backpropagation algorithm *each time* a sentence constituent is heard. At best, the burden of proof rests upon St.J&Mc to show that this application of the backpropagation algorithm has even a *rough* physiological basis.

Of course, St.J&Mc are by no means alone in their questionable use of backpropagation. A glance at recent proceedings of the Cognitive Science Society reveals many suspect uses of backpropagation. Often the algorithm is employed when no plausible source of error correction exists. Most of the systems discussed herein employ backpropagation without offering more than a slender excuse for doing so. This is not surprising;

the undefended use of backpropagation among connectionists has become so common that one hardly notices that justifications are seldom offered. Again, the situation is strongly analogous to that which has existed in classical AI. Eventually people cease to question whether increasingly convoluted algorithms have any biological foundation whatever.

Elman (1989, 1990)

Like St.J&Mc, Elman employs backpropagation, but he does attempt a justification for doing so. Recall that, in an effort to teach his networks the syntactic categories of lexical items, Elman trains the network, via backpropagation, to *attempt* to predict the *next* word in a successive stream of words. In defense of this "error-feedback" strategy, Elman remarks that "it does seem to be the case that much of what listeners do involves anticipation of future input". Presumably, Elman takes this as evidence that listeners are constantly attempting to predict the next input element. Within my own experience, I find little introspective evidence to support this *strong* claim. However, even if we ignore introspective evidence, a more serious objection to Elman's remark arises once we reflect that, until listeners have acquired a degree of syntactic knowledge, they have no basis upon which to found predictions of succeeding input. It is thus implausible that they would attempt to do so. Moreover, Elman's invocation of backpropagation after *each word* is processed is subject to the same criticism as St.J&Mc's usage. It is difficult to believe that a biological process supporting anything analogous to backpropagation could occur between succeeding words in an utternace.

Another difficulty with Elman's approach is that, when relative clauses are involved, training occurs in 4 distinct phases. Phase 1 presents the network with a concatenated string of 10,000 *simple* grammatical sentences (no relative clauses are included). This string of 10,000 sentences is presented to the network 5 times over. Now, not by the wildest stretch of the imagination is this a psychologically plausible regime. Normally, a child would encounter many breaks even during a series of 20 sentences. During some of these breaks the child may hear sentence fragments, or even simple names. Almost certainly, the child would be exposed to a substantial percentage of unfinished and ungrammatical sentences. The question naturally arises, would Elman's networks be able to induce systematic regularities under these conditions? Not likely, but if not, what are the real implications of this research?

Returning to the succeeding phases of Elman's regime, phase 2 modifies phase 1 by having 25% of the 10,000 sentences contain relative clauses. Phase 3 contains 50% relative clause sentences, and phase 4 contains 75% relative clauses. Clearly, this training regime is highly contrived. Children are not exposed to anything like this artificial partitioning of the input copora.

In fairness, I should note that the artificiality of Elman's training regime is certainly not unique to his work, and he would no doubt readily concede its artifice. Somewhat analogous remarks would apply to St.J&Mc (who also train their c-nets with implausibly long strings of sentences), to Chalmers, and to Smart. It is doubtful whether any of these authors would attempt serious defenses of the size and presentation of their input copora. However, in the absence of such defense we must ask whether these networks could discover even the moderate degree of compositionality they do discover if they were subjected to the erratic, mixed, and often ungrammatical input that humans receive.

References

Chalmers, D. (1990) "Why Fodor and Pylyshyn Were Wrong: the Simplest Refutation", Proceedings of the Twelfth Annual Conference of the Cognitive Science Society, Cambridge, Mass.

Dromi, E. (1987) *Early Lexical Development*, Cambridge University Press, Cambridge, UK.

Elman, J.L. (1989) "Representation and Structure in Connectionist Models", CRL Technical Report 8903, University of California at San Diego.

Elman, J.L. (1990) "Finding Structure in Time", *Cognitive Science*, Vol. 14, pp. 179-212.

Fodor, J.A. (1976) *The Language of Thought*, Harvard University Press, Cambridge, Mass.

Fodor, J.A. & Pylyshyn, Z.W. (1988) "Connectionism and Cognitive Architecture: A Critical Analysis", *Cognition*, Vol. 28, pp. 3-71.

Hadley, R.F. (1990) "Connectionism, Rule Following, and Symbolic Manipulation", Proceedings of (AAAI) the American Association for Artificial Intelligence, MIT Press, Boston.

Ingram, D. (1989) *First Language Acquisition*, Cambridge University Press, Cambridge, UK.

Katz, N., Baker, E. & Macnamara, J. (1974) "What's in a Name? A Study of Children Learn Common and Proper Names", *Child Development*, Vol. 45, pp. 469-473.

Pollack, J.B. (1990) "Recursive Distributed Representations", *Artificial Intelligence*, Vol. 46, pp. 77-105.

Rumelhart, D.E., Hinton, G., & Williams, R. (1986) "Learning Internal Representations through Error Propagation", in D.E. Rumelhart, J.L McClelland and the PDP Research Group (Eds.), *Parallel Distributed Processing; Experiments in the Microstructure of Cognition 1: Foundations*, MIT Press, Cambridge, MA.

Small, S.L. (1990) "Learning Lexical Knowledge in Context: Experiments with Recurrent Feed Forward Networks", Proceedings of the Twelfth Annual Conference of the Cognitive Science Society, Cambridge, Mass.

Smolensky, P. (1988) "On the Proper Treatment of Connectionism", *Behavioral and Brain Sciences*, Vol. 11, pp. 1-23.

Smolensky, P. (1990), "Tensor Product Variable Binding and the Representation of Symbolic Structures in Connectionist Systems", *Artificial Intelligence*, Vol. 46, pp. 159-216.

St. John, M.F. & McClelland, J.L. (1990) "Learning and Applying Contextual Constraints in Sentence Comprehension", *Artificial Intelligence*, Vol. 46, pp. 217-257.

Centre for Systems Science

• Simon Fraser University established the **Centre for Systems Science** as a provincial centre for advanced study. The objective of CSS is twofold: coordinate and combine advanced systems research at SFU, and encourage application of research results within industry in order to further the technological development of British Columbia.

• CSS has assembled three research groups to streamline its efforts. The orientation of these groups is as follows: intelligent systems, microelectronics, and computer and communications systems. Because of shared research interests, this latter group merges with LCCR (**Laboratory for Computer and Communications Research**) for the publication of technical reports.

• CSS's three groups involve over 80 skilled faculty researchers principally from SFU's Schools of Engineering Science and Computing Science, but also draw substantially on Mathematics, Physics, Geography, Philosophy, Linguistics, Psychology, Communication, and Kinesiology.